research articles Journal of proteome research

A "Tagless" Strategy for Identification of Stable Protein Complexes Genome-wide by Multidimensional Orthogonal Chromatographic Separation and iTRAQ Reagent Tracking

Ming Dong,[†,⊥] Lee Lisheng Yang,[†,⊥] Katherine Williams,[‡] Susan J. Fisher,[†,§,∥,⊥] Steven C. Hall,[§,∥,⊥] Mark D. Biggin,[†,⊥] Jian Jin,[†,⊥] and H. Ewa Witkowska*[,§,∥,⊥]

Lawrence Berkeley National Laboratory, Berkeley, California 94720, Applied Biosystems, Foster City, California 94404, UCSF Mass Spectrometry Core Facility and Department of Cell and Tissue Biology, University of California San Francisco, San Francisco, California 94143, and Virtual Institute for Microbial Stress and Survival, Berkeley, California 94720

Received September 26, 2007

Tandem affinity purification is the principal method for purifying and identifying stable protein complexes system-wide in whole cells. Although highly effective, this approach is laborious and impractical in organisms where genetic manipulation is not possible. Here, we propose a novel "tagless" strategy that combines multidimensional separation of endogenous complexes with mass spectrometric monitoring of their composition. In this procedure, putative protein complexes are identified based on the comigration of collections of polypeptides through multiple orthogonal separation steps. We present proof-of-principle evidence for the feasibility of key aspects of this strategy. A majority of *Escherichia coli* proteins are shown to remain in stable complexes during fractionation of a crude extract through three chromatographic steps. We also demonstrate that iTRAQ reagent-based tracking can quantify relative migration of polypeptides through chromatographic separation media. LC MALDI MS and MS/MS analysis of the iTRAQ-labeled peptides gave reliable relative quantification of 37 components of 13 known *E. coli* complexes: 95% of known complex components closely co-eluted and 57% were automatically grouped by a prototype computational clustering method. With further technological improvements in each step, we believe this strategy will dramatically improve the efficiency of the purification and identification of protein complexes in cells.

Keywords: tagless strategy • protein complex • protein separation • column chromatography • relative quantitation • LC MALDI workflow • MALDI TOF/TOF • iTRAQ reagent • Pearson cluster analysis • *E. coli*

Introduction

Homomeric and heteromeric protein complexes are distinctly shaped, highly organized, and often specifically localized "molecular machines." It is these complexes, rather than single polypeptides, that are the elemental components of functioning biological systems, and thus, characterizing them is a prerequisite for any informed intervention aimed at modifying cellular processes. In recognition of this need, the U.S. Department of Energy's Genomics:GTL program, of which this work is a part, has established as a major goal the development of very high-throughput methods to characterize the structures and functions of protein complexes in microbes relevant to its mission.[1]

Historically, stable protein complexes were identified one at a time, often as the result of purifying an enzyme activity of interest. In this traditional approach, complexes were inferred when multiple polypeptides comigrated together with an associated enzyme activity through multiple chromatographic separation steps,[2–4] demonstrating the same sedimentation velocities[5,6] or electrophoretic mobilities. [7] More recently, stable protein complexes have been identified using high-throughput mass spectrometric detection of collections of polypeptides that are stably associated with heterologous affinity-tagged polypeptides. [8] In particular, tandem affinity purification[9–12] has proven to be highly effective in mapping the soluble portion of the yeast[13,14] and *Escherichia coli*[15] interactomes. Despite its undoubted utility, however, tandem affinity purification (TAP) suffers from several limitations. For example, this method is restricted to biological systems that are amenable to the genetic manipulations required to introduce the affinity-tagged polypeptides into cells. Furthermore, the addition of an affinity tag may destabilize some protein–protein interactions or alter other relevant protein activities. Finally, TAP requires a distinct genetic strain to be constructed for each polypeptide and then each strain must be separately cultured and analyzed. These

* To whom correspondence should be addressed at 521 Parnassus Ave., Box 0512, San Francisco, CA 94143. Tel. office: (415) 502-4019. Tel. lab: (415) 502-7502. E-mail: witkowsk@cgl.ucsf.edu.
† Lawrence Berkeley National Laboratory.
⊥ Virtual Institute for Microbial Stress and Survival.
‡ Applied Biosystems.
§ UCSF Mass Spectrometry Core Facility, University of California San Francisco.
∥ Department of Cell and Tissue Biology, University of California San Francisco.

10.1021/pr700624e CCC: $40.75
Published on Web 03/13/2008

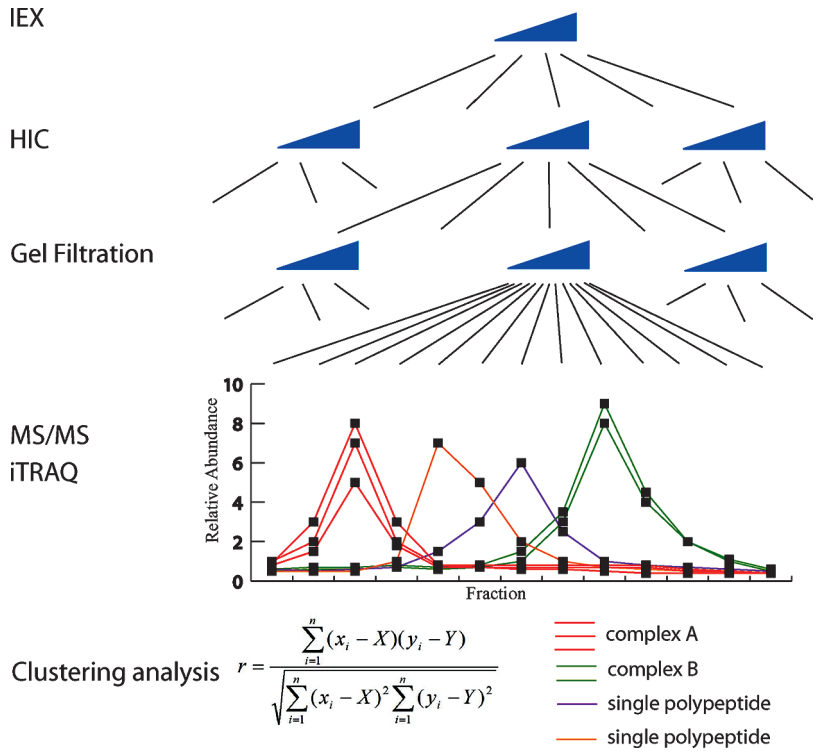

Figure 1. The concept of the "tagless" protein complex identification strategy. A crude cell lysate is fractionated successively by highly parallel, orthogonal purification steps: in the example given, ion exchange (IEX), hydrophobic interaction (HIC), and gel filtration chromatography. A rational sampling of fractions from the preceding separation step is submitted to the subsequent separation step, generating thousands of fractions at the last purification step. Selected fractions from the last step are then subjected to proteolytic digestion and iTRAQ reagent labeling, and the products are then analyzed by mass spectrometry to identify polypeptides and measure their relative abundances as they migrate through the separation media; the iTRAQ reagent serves as a quantitative beacon of protein presence. Similarities among polypeptide elution profiles are evaluated using clustering analysis. Putative complexes are defined as sets of polypeptides that cluster at an experimentally established confidence level. In the example shown, two putative heteromeric complexes A and B, composed of three and two co-eluting components, respectively, are shown. Two proteins with no co-eluting partners are also detected. Since the last separation step is based on molecular weight, it can be determined if these noncoeluting polypeptides are either monomers or homomeric complexes. Note that the pilot experiments described in the paper do not use the exact series of fractionation steps shown in this concept figure.

and other limitations suggest that it may prove difficult to automate this strategy to achieve higher throughput than has been already attained.

To overcome these limitations, we are developing a novel method to rapidly purify and identify the majority of stable protein complexes in a cell without the use of affinity tags or affinity purifications.[16] This "tagless" strategy detects polypeptides in endogenous complexes isolated from wild-type cells based on the shared elution profiles of polypeptides that, as components of a protein complex entity, comigrate through multiple chromatographic steps. A diagrammatic representation of the experimental strategy is shown in Figure 1. Starting with a crude extract prepared from a single large culture of cells, a number of orthogonal chromatographic separation steps are performed under conditions that preserve interactions among the polypeptide components of the complexes. Selected fractions from each column are used as the input for the next step. *In toto*, several hundred parallel chromatography runs are performed, generating thousands of fractions that are then proteolyzed prior to labeling of the peptide products with iTRAQ reagents.[17] Matrix-assisted laser desorption ionization time-of-flight (MALDI TOF) tandem mass spectrometry (MS/MS) quantification of the relative abundance of each polypeptide across sets of fractions from each column generates elution profiles for all detected proteins. The migration of polypeptides through earlier chromatographic steps can be inferred from MS analysis of the final fractions since this experimental strategy captures the complete repertoire of proteins from each column. Heteromeric complexes are inferred when clustering analysis of protein elution profiles reveals subsets of polypeptides that have similar elution patterns. Homomeric complexes are inferred when single polypeptides that do not co-elute with other polypeptides migrate in either size exclusion chromatography or native PAGE with an apparent molecular weight (MW) significantly larger than predicted from genomic sequence data. A similar concept named "protein correlation profiling" that utilizes mass spectrometry as a tool of monitoring protein comigration[18] was introduced by Matthias Mann to identify a subpopulation of centrosomal proteins separated *via* sucrose gradient centrifugation. Intrinsic to both this and the tagless strategy scenarios, purification of complexes to homogeneity is not required to reliably classify polypeptides as members of the same complex.

In this work, as a proof-of-principle, we present data from experiments in which an *E. coli* lysate was passed through a pilot tagless pipeline with the intent of answering four key questions about this approach. (i) Are protein complexes stable through multiple orthogonal chromatographic separation steps? (ii) Is column chromatography sufficiently reproducible? (iii) Is an iTRAQ-based MS method sufficiently accurate and

reproducible to enable the relative quantitation of comigrating polypeptides in neighboring and otherwise related fractions? (iv) Is clustering analysis capable of automatic detection of complexes from the resulting mass spectrometry data? In all four cases, the experimental data supported the feasibility of the method we propose. The findings have important relevance to large-scale biology projects. Compared to TAP, this new approach offers several advantages. *For example*, just as shotgun approaches have greatly accelerated genome-wide sequencing projects, a tagless method, which is highly automatable, could allow much higher throughput characterization of protein complexes at an organismal level.

Materials and Methods

Protein Separation. *E. coli* lysates and protein extracts (20–50 mg/mL) were prepared as previously described.[19] All separations were performed at 4 °C except for hydrophobic interaction chromatography (HIC), which was performed at room temperature. All chromatographic columns were run using an FPLC (GE Healthcare), and protein elution was monitored at 280 nm. Chromatographic fractions were analyzed by SDS-PAGE using the Criterion Precast gel system (Bio-Rad): 4–15% SDS-PAGE gradient gels were used for SDS-PAGE and 4–20% Native PAGE gradient gels for Native PAGE. Gels were stained using a SilverQuest silver staining kit (Invitrogen). Protein concentrations were measured by Bradford assay (Pierece).

1. Protein Separation for Studies of Protein Complex Stability during Chromatography. Ten to 15 mg of *E. coli* protein extract was diluted 2-fold with Buffer A (25 mM HEPES, 10% glycerol, 0.01% NP-40, and 2 mM DTT) and applied to an 8 mL Mono Q column (GE Healthcare) equilibrated with buffer A containing 100 mM NaCl. The column was developed with a 100−600 mM NaCl gradient in buffer A (25 column volumes, cv) with collection of 2 mL fractions (25% cv) using a flow rate of 2 mL/min. One of the eluted Mono Q column fractions was further separated by gel filtration using a 23.5 mL Superose 6 column (GE Healthcare) equilibrated with buffer A containing 100 mM NaCl and eluted with the same buffer with a flow rate of 0.2 mL/min.

2. Protein Separation for iTRAQ Quantification and Protein Complex Purification. The protein separation for iTRAQ analysis was performed at two different scales using gel filtration chromatography followed by anion exchange chromatography. Protein extract (50 or 500 mg) was loaded onto a 120 or 320 mL Sephacryl S-400 column (GE Healthcare) equilibrated with buffer A containing 100 mM NaCl, and the column was then developed with the same buffer at a flow rate of either 0.3 or 1 mL/min for the small- and large-column, respectively. Two major UV peaks were observed in both cases: the earlier eluting peak contained high molecular weight (HMW) proteins (above 200 kDa) that constituted $^1/_7$ to $^1/_{10}$ of the total eluted protein. The HMW protein fractions derived from the small- and large-scale preparations were separated using an 8 or 20 mL Mono Q column that was developed with a 25 cv linear gradient from 100 to 600 mM NaCl in buffer A at a flow rate of either 2 or 4 mL/min. The 25% cv or 10% cv fractions were collected from the small- and large-scale purifications, respectively.

To purify particular protein complexes, selected Mono Q fractions were further fractionated by either gel filtration on a 23.5 mL Superose 6 column (for pyruvate dehydrogenase) or by hydrophobic interaction chromatography on a 1.7 mL Source 15PHE column (GE Healthcare) (for RNA polymerase and 60 kDa chaperonin). Superose 6 columns were run as described in the previous section. Hydrophobic interaction columns were equilibrated with buffer B (25 mM HEPES, 10% glycerol, and 2 mM DTT) containing 1 M $(NH_4)_2SO_4$, and the column was developed with a linear gradient from 1 to 0 M $(NH_4)_2SO_4$ in buffer B. Protein fractions from the Superose 6 or hydrophobic interaction columns were analyzed by SDS-PAGE and were not monitored by iTRAQ.

Protein Digestion and Labeling with iTRAQ Reagents. Selected portions of the anion exchange chromatography eluates were sampled for mass spectrometry analyses at a frequency of 25% or 50% cv. Specifically, 1 in 2 or 1 in 6 fractions were assayed, a total of 7 and 15 fractions for the small- and large-scale experiments, respectively. The protein content of the fractions was estimated by using the Bradford assay.[20] This information was used to ensure that protein digestion and derivatization for each experiment were performed at similar protein concentrations. Equal fraction volumes were digested and labeled when their respective protein concentrations were within 100% of each other. Otherwise, fraction volumes with equal amount of protein were used as the starting material. Briefly, the proteins in each fraction were precipitated with acetone (6× volume excess), solubilized in 100 mM triethylammonium bicarbonate buffer (TEAB, pH 8.5) containing 0.1% SDS, reduced with Tris-(2-carboxyethyl) phosphine (TCEP), alkylated with methyl methanethiosulfonate (MMTS), and digested with porcine trypsin (Pierce) at 37 °C overnight. The resulting tryptic peptide mixtures were derivatized with iTRAQ reagents in the TEAB buffer/80% ethanol for 1 h at room temperature. The manufacturer's protocol for iTRAQ reagent labeling was followed; however, an approximate 4–5× higher iTRAQ reagent/protein ratio was used at the protein scale of ∼20–25 μg. Postlabeling, four consecutive Mono Q fractions, each tagged with a different iTRAQ reagent, were combined to generate a multiplexed sample; consecutive multiplexed samples shared one common fraction. The sample volume was reduced to ∼10–20 μL on a SpeedVac prior to one-step cation exchange chromatography which was carried out using the resin-containing cartridge and buffers provided by the manufacturer.[17] The elutes that contained the peptide mixtures were concentrated to a volume of 10–20 μL and stored at −20 °C prior to MALDI LC MS/MS analysis.

LC MALDI MS/MS. A Pepmap C18 trap column and a nanocolumn (100 μm i.d., 15 cm length, Dionex/LC Packings) were used for desalting and reversed-phase (RP) peptide separation, respectively. A 30 min linear gradient from 2% B to 40% B was run at 500 nL/min flow rate, utilizing solvents A, 2% AcCN/0.1% trifluoroacetic acid (TFA), and B, 85% ACN/5% isopropanol, 1.0% TFA using an Ultimate LC System (Dionex/LC Packings). Reversed-phase-separated peptides were collected directly onto a stainless steel MALDI target utilizing Probot (Dionex/LC Packings) spotting robot. Column elute was combined, in a mixing tee, with MALDI matrix (α-cyano-4-hydroxycinnamic acid, 6 mg/mL in 80% ACN/0.1% TFA/10 mM dibasic ammonium phosphate), containing 25 fmol/μL Glu-fibrinopeptide (GluFib) for internal calibration, delivered at 1 μL/min. Peptides were analyzed on a 4700 and 4800 Proteomics Analyzer mass spectrometer (Applied Biosystems/MDS Sciex) in the positive ion mode. The 4700 and 4800 Proteomics Analyzers were equipped with TOF/TOF ion optics and a 200 Hz Nd:Yag laser.[21] For collision-induced dissociation (CID), the collision cell was floated at 1 kV (4700) or 2 kV (4800), the resolution of the precursor ion selection was set to 200 and

300 fwhm for the 4700 and 4800 analyzers, respectively, and air was used as the collision gas at 5×10^{-7} Torr. Automated acquisition of MS and MS/MS data was controlled by 4000 Series Explorer Software. Internal one-point calibration utilized m/z of monoisotopic molecular ion of GluFib that met the following acceptance criteria: S/N 50, mass error 50 ppm; when the acceptance criteria were not met, default calibration based on a plate model algorithm (Applied Biosystems) was employed. Typical mass accuracy was within 10 and 50 ppm for the internal and default calibration, respectively. Automated MS/MS data analysis was performed utilizing GPS Explorer software 3.5 with MASCOT 2.1.0 (Matrix Science) software for protein identification and quantitation of iTRAQ reporter ions. The following criteria were employed for generation of MS/MS peak list: S/N 5, m/z 50 to −20 from a precursor molecular ion, 50 peak limit per 200 Da, a maximum number of peaks 80. *E. coli* taxonomy within Swiss-Prot protein database, release 48.0 of 13-September-2005 and release 49.6 of 02-May-2006, was interrogated for the data sets generated on 4700 for all 15 fractions and on 4800 for the first 10 fractions, respectively. The following search parameters were utilized: precursor mass tolerance, 50 ppm; fragment mass tolerance, 0.15 Da; tryptic digestion with 2 missed cleavages; fixed modifications, S-MMTS, K- iTRAQ, and N-term iTRAQ; variable modifications, deamidation (Asn and Gln); Met-sulfoxide. GPS Confidence Interval (C.I. %) of 95% was used as the acceptance criteria, and hence, identification of each polypeptide was based upon at least one peptide that scored above a threshold value set by the Mascot search engine to indicate identity or extensive homology of proposed sequence at $p < 0.05$. The reported protein list was manually updated to reflect the UniProt protein entry names and accession numbers (release 53.2 of 26-June-2007); EcoCyc database[22] (http://ecocyc.org/) was utilized to facilitate this process. Average relative ratios were calculated for each polypeptide using the GPS Explorer 3.5 algorithm without invoking a "bias" correction option. Only peptides that were completely labeled with iTRAQ at N-termini and lysines and whose individual relative ratios were different from zero were considered while calculating protein average. The outliers were automatically excluded. Finally, the results for each multiplex were normalized to represent the same volume of the column fraction taken for the analysis.

Evaluation of Quality of Quantitation Data and Biological Reliability of Measurements. To evaluate the extent of side reactions, the data were reanalyzed by interrogating the same database and using the same parameters as described above with the exception of iTRAQ settings, this time specifiying a flexible rather than a fixed modification type and allowing for tyrosine derivatization. Only a limited number of under-derivatized peptides was revealed and no hits carrying iTRAQ-labeled tyrosine were found. To minimize the number of overlapping precursors, precursor ion selection for MS/MS data acqusition was performed at the resolution as high as possible without significantly jeopardizing sensitivity, and a filter of a minimum of 200 resolution between a target precursor and potential nonrelated molecular ions was applied. Nevertheless, given the complexity of the sample and the limitation of the TOF/TOF precursor ion selection window, it is inevitable that some of the quantitation data might have been adversely affected by interfering ions. A potential presence of multiple precursors was not addressed by the GPS software, and no systematic examination of all the data was undertaken to evaluate the extent of the possible problem. However, a limited number of MS and MS/MS spectra, predominantly those derived from proteins representated by a small number of peptides, were examined manually, and in the great majority of cases, no significant level of unexplained (product ion) signals was observed (Supporting Information, Supplementary Figure 3). We have also examined a number of outliers among multiple peptides representing abundant proteins and only a small number of them could be possibly explained by the presence of a detectable interfering ion close to the intended precursor (data not shown). Reproducibility of the methods described in this study is presented and discussed in Results and Discussion. The aim of the study was to compare the results of the tagless strategy with the known information on protein complexes and protein–protein interactions in the model organism, *E. coli*. It was not intended to be a discovery study, and hence, no attempt was made to validate any unknown, putative, complexes that might have been detected using a clustering algorithm (Supporting Information, Supplementary Table C)

Deriving Polypeptide Elution Profiles. A graphic representation of the process of generating a polypeptide elution profile is shown in Supplementary Figure 1 in Supporting Information. In the first step, the polypeptide elution profile was generated within each multiplex (Supplementary Figure 1, panel A, in Supporting Information). In the second step, the independent partial elution profiles that were generated in the previous step were colapsed into a single elution profile across the whole chromatogram by equalizing the values in the shared fractions of each consecutive multiplex pair, starting from the beginning of the chromatogram (Supplementary Figure 1, panel B, in Supporting Information). Finally, in the third step, the relative polypeptide abundance was determined by arbitrarily assigning a value of 1.0 to the apex of a polypeptide elution peak and scaling all the other data points accordingly (Supplementary Figure 1, panel C, in Supporting Information). We point out that the actual values of relative abundances of different polypeptides across the chromatogram cannot be directly compared since they are independently derived from different sets of polypeptide-specific peptides.

Polypeptide Clustering. To identify putative protein complexes, a comparison of polypeptide elution profiles was performed within all the fractions where the polypeptide was observed. Average relative ratios calculated for each polypeptide by the GPS Explorer 3.5 algorithm that were normalized and scaled, as described above, were employed for clustering analysis. The first step was to identify all valid profile peaks using the following process: (i) find the center, left, and right edges for all elution peaks for all polypeptides using a simple peak detection algorithm developed in our laboratory and (ii) filter out the noise. The latter was accomplished by examination of the peak intensity ratios relative to the highest peak in the same polypeptide profile (R1) and relative to the intensities of its own left and right edges (R2). If any of the ratios, R1 and/or R2, were below the threshold ($R1 \leq 0.15$ and $R2 \leq 1.20$), the peak was classified as noise. The R1 and R2 threshold values are dependent on the data complexity and quality and might need further tuning in the future as the data size grows. Once a set of elution peaks of all polypeptides was established, Pearson correlation coefficients between any two peaks that overlap significantly were calculated. In this work, Pearson correlation coefficients were used as a measure of similarity between two peaks, see the formula below where (x_1, x_2, x_3, . . . x_n) and (y_1, y_2, y_3, . . . y_n) are normalized intensities of

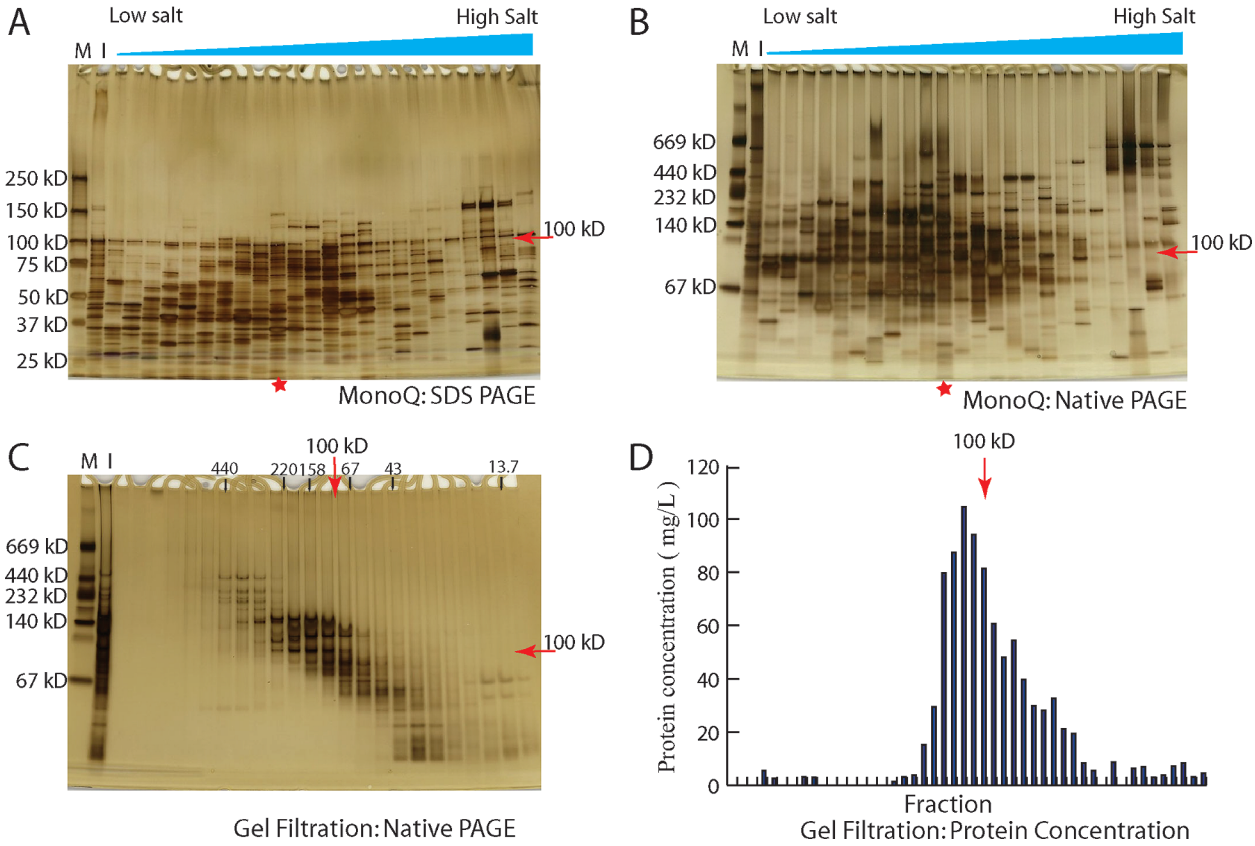

Figure 2. Typical protein size distribution in partially fractionated *E. coli* lysates. The results of the SDS-PAGE (4–15%) and native PAGE (4–20%) analyses of selected fractions of a crude lysate after separation by Mono Q chromatography are shown in panels A and B, respectively. Panel C shows the results of the Native PAGE (4–20%) analysis of fractions collected following gel filtration chromatography (Superdex 200 column) of one of the Mono Q column fractions, annotated with an asterisk in Panels A and B. Panel D shows the distribution of total protein concentration across the Superdex 200 column; the red arrow marks the estimated position of elution of a 100 kDa species. Both Native PAGE and gel filtration data suggest that the majority of protein, by mass, participates in complexes. A comparison of Native PAGE of the protein fraction annotated with an asterisk in panel B and products of its further separation (panel C) shows that the proportion of high versus low molecular weight species changes little during size exclusion chromatography, indicating that there is only slight dissociation of complexes during this step.

peaks x and y across fractions 1, 2, . . ., n, and X and Y are their average intensities across all n fractions, respectively.

$$r = \frac{\sum_{i=1}^{n}(x_i - X)(y_i - Y)}{\sqrt{\sum_{i=1}^{n}(x_i - X)^2 \sum_{i=1}^{n}(y_i - Y)^2}} \quad (1)$$

The clustering analysis routine was based on an algorithm originally developed for evaluation of gene expression profiles (http://genetics.stanford.edu/sherlock/cluster.html). This algorithm was customized to accommodate our polypeptide elution profile data. Mathematical averages of coefficient values of clustered peaks were used as the metrics for similarity measurement. On the basis of these criteria, a putative complex is called if the average Pearson coefficient of a cluster of polypeptides exceeds a threshold value of 0.92.

Results and Discussion

Do Protein Complexes Survive Multiple Chromatographic Separation Steps? A large body of evidence from biochemical analyses performed over the last several decades suggests that many protein complexes are stable under the conditions that are used in typical chromatographic separations (e.g., McHenry and Crow,[2] Srere and Mathews,[3] Austin and Biggin[4]). At the same time, low affinity protein–protein interactions with micromolar dissociation constants are not expected to survive multiple separation techniques. To roughly estimate the proportion of cellular polypeptides that are engaged in interactions sufficiently stable to be detected by the tagless strategy, proteins from an *E. coli* crude extract were first fractionated by chromatography over a Mono Q anion exchange column. SDS-PAGE of the proteins eluting in each fraction showed that the estimated molecular weights of the large majority of polypeptides were less than 100 kDa (Figure 2A). In contrast, native PAGE separation of the same fractions suggested that approximately half of the proteins by mass had a M_r greater than 100 kDa (Figure 2B). Since the migration of proteins in native gels is a function of both their molecular weight and net charge, we further separated several Mono Q fractions by gel filtration chromatography as this charge-independent fractionation method depends only on molecular weight and shape. Native PAGE of the resulting fractions showed that there was a broad correlation between the mobilities of proteins in native PAGE (e.g., Figure 2C) and their molecular weights as estimated by gel filtration chromatography (Figure 2D). In a typical experiment, 56% of the protein mass eluted from the gel filtration column with an apparent molecular weight greater than 100

kDa (Figure 2D), whereas only a few polypeptides greater than 100 kDa were observed in SDS-PAGE of the material loaded onto a column (Figure 2A, asterisked lane). Thus, both native gel electrophoresis and gel filtration suggested that a majority of *E. coli* proteins by mass were found in complexes with estimated molecular weights that were greater than the individual polypeptides that constitute the organism's proteome. While this analysis does not estimate the fraction of polypeptides that formed protein complexes with MW less than 100 kDa, we suggest that a similarly high proportion of proteins in this molecular weight range are likely to exist as complexes.

In addition, a comparison by native PAGE of protein fractions before and after the second chromatographic step showed few changes in the proportion of high- versus low-molecular weight species during size exclusion chromatography, indicating that a majority of protein complexes were stable during this step (compare Figure 2C with 2A, lane marked with an asterisk). Subsequent analyses using the same approach suggested that a majority of stable protein–protein interactions were not disrupted by hydrophobic interaction chromatography or ammonium sulfate fractionation (data not shown).

Approximately 40% of *E. coli* polypeptides form complexes that are sufficiently stable that they can be detected in reciprocal TAP isolation experiments.[15] This percentage is very similar to our estimate of the fraction of polypeptides that were engaged in stable protein–protein interactions following the various separation methods that were used in our pilot experiments. TAP and the techniques we employed take similar amounts of time to separate complexes from a crude lysate. Thus, it seemed likely that there will be a large overlap in the protein complexes isolated by these two approaches, an idea that was supported by further experiments that are described below.

Choice of iTRAQ-Based MALDI MS/MS for Protein Elution Profiling. A major challenge in establishing the feasibility of our proposed tagless strategy was to select a suitable mass spectrometry method. Because of the large number of fractions to be analyzed, it was critical to adopt an approach that minimized the number of MS/MS analyses as this could otherwise become a serious rate-limiting step. It was also essential to adopt a method that was able to quantitate relative abundances of polypeptides in different fractions.

We chose a LC MALDI MS workflow, rather than a LC electrospray ionization (ESI) workflow, as this decouples the LC step from the MS and MS/MS steps and, thus, allows repeated interrogation of archived MALDI sample plates. In the context of our proposed analysis of a series of closely related fractions of similar content (Figure 1), information generated in the course of MS/MS runs on preceding fractions can then be used to design more efficient MS/MS data acquisition strategies for the fractions that follow, thus, reducing the overall time of MS/MS analysis.

To track changing relative abundances of polypeptides between fractions, we chose an isotopic dilution method that employs the primary amine-directed iTRAQ reagent as the label.[17,23] The iTRAQ labeling methodology is the most robust high-throughput means of quantifying protein relative abundances by MALDI TOF MS/MS and offers an accuracy and precision comparable with the label-free ESI-based[24–27] and MALDI-based[28] methods. Furthermore, unlike other potential mass spectrometry labeling methods,[29] iTRAQ multiplexes four samples in one analysis, further reducing the number of MS/MS analyses required.

The iTRAQ reagent was originally developed for comparing relative levels of peptides in protein expression profiling experiments. We have adapted it in the following way for our purposes (Figure 3). A subset of fractions from a single chromatographic separation step is analyzed at a frequency based on the resolution of the chromatography method. For example, every other fraction across a column may be selected. Small aliquots from these fractions are then digested with trypsin and labeled with one of the four different iTRAQ reagents such that each fraction within a multiplex set is represented by one of the four iTRAQ reporter ions at m/z of 114, 115, 116, or 117. Each set contains four sequential fractions from the protein column elute. Every pair of adjacent multiplexes shares a border fraction, and hence, all multiplexes representing a single protein separation step are strung together *via* a series of shared fractions (Figure 3). LC MALDI TOF MS/MS analysis of each multiplex set produces sequence-specific, gas-phase product ions from which a peptide is matched to its parent polypeptide. Concurrently, the parent polypeptide's relative abundance, within the analyzed fraction set, is calculated based on the intensities of four iTRAQ reporter ions. A similar approach of using either iCAT or iTRAQ reagent-labeling to follow protein gradient distribution profiles under conditions of sedimentation was recently introduced by Kathryn Lilley et al.[30–32]

A Proof-of-Principle Demonstration. To provide a first proof-of-principle that iTRAQ quantification allows protein complexes to be detected, we first performed a minimal two-step chromatographic separation of an *E. coli* crude extract, which takes high molecular weight proteins derived from a size exclusion column and fractionates them further by Mono Q anion exchange chromatography. We then quantitated the relative levels of five RNA polymerase subunits across the Mono Q fractions.[16] In this and subsequent iTRAQ analyses, fractions were sampled at a frequency such that they were separated by at least one fraction and by no more that 25−50% of a column volume as this was found to provide sufficient resolution to detect comigration of polypeptides belonging to known complexes. The fractions themselves were quite heterogeneous, being derived from only two chromatography steps, and contained a broad mixture of many polypeptides (Figure 4A). Nonetheless, despite this crude fractionation, between 6 and 30 tryptic peptides were detected for the five known subunits of RNA polymerase. The iTRAQ quantification showed that the individual peptides derived from a given polypeptide gave similar, albeit, not overlapping relative concentration profiles across the fractions (Figure 4B). We also observed similar variation between peptides in model studies on standard proteins (unpublished data), suggesting that this variation resulted from differential rates of peptide generation during tryptic digestion and/or losses during sample processing, rather than being indicative of structural heterogeneity within the proteins present in each fraction. While relative abundances of some tryptic peptides varied significantly from the mean, elution profiles based upon their relative ratios differed from the average elution profile in amplitude but not in localization of apexes (Figure 4B). Hence, if only few peptides are detected due to low polypeptide abundance, it is desirable to monitor the same set of tryptic peptides representing the polypeptide of interest in all fractions that are analyzed. When a sufficient number of peptides is detected, the mean data for multiple peptides should be the best guide to the relative abundance of each polypeptide. Indeed the averaged profiles for all five

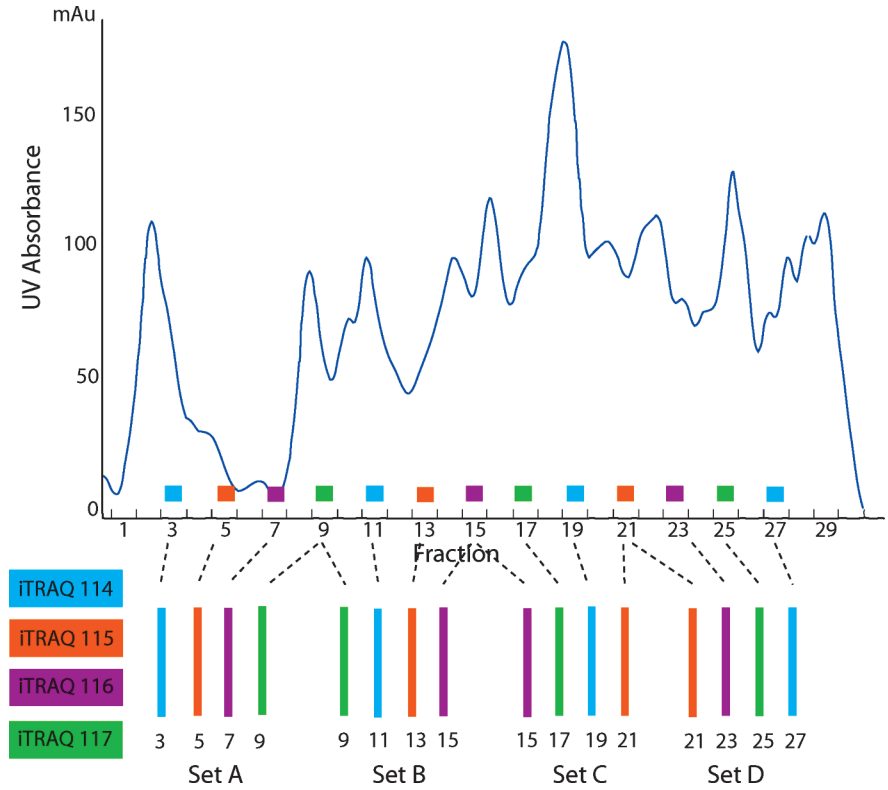

Figure 3. The experimental workflow of protein identification and quantification. Eluted proteins are sampled at a frequency dependent upon the resolution of the separation step. In the example shown, every other fraction is sampled. The appropriate volume of each fraction is withdrawn so that each fraction is represented in the final four-plex set by the same amount of total protein (~20 μg). Proteins from each fraction are independently digested with trypsin and labeled with iTRAQ reagent. Four successive fractions, each labeled with a different iTRAQ reagent, are combined to form multiplexes, annotated as A, B, C, and D at the bottom of the figure. Each pair of adjacent multiplexes shares one bordering fraction. Each multiplex is analyzed by a reversed-phase nanoLC MALDI MS and MS/MS. Elution profiles are generated for each detected protein on the basis of iTRAQ-derived relative abundances within all multiplexes, as described in Materials and Methods.

polymerase components were very similar to each other (Figure 4C), consistent with the known tight association of these five polypeptides.

The above results suggest that iTRAQ quantitation is sufficiently accurate to detect comigrating complex components. Since the fractions analyzed contain far more proteins than the highly purified fractions envisioned being assayed in our finalized tagless strategy protocol (Figure 1), the fact that the iTRAQ-based method was effective in these less than optimal circumstances was encouraging.

Reproducibility of iTRAQ-Based Protein Elution Profiling. In spite of this encouraging result, if a full-scale implementation of the tagless strategy is to be successful, iTRAQ-based quantitation and column chromatography will have to be sufficiently reproducible so that data from different fractions, multiplex sets, columns, and days can be compared as part of a large single data set. Therefore, reproducibility of the tagless method was examined at three levels: (a) reproducibility of mass spectrometric data acquired on a single instrument with the same spotted samples; (b) reproducibility of tryptic digestion, labeling, and other sample preparation steps; and (c) reproducibility of replica chromatography separations of protein mixtures.

Repeated analysis of the same LC MALDI plate gave essentially the same iTRAQ ratio values for relative abundances of polypeptides, indicating high analytical reproducibility of the mass spectrometers employed in this study (data not shown). Duplicate proteolytic digestion and iTRAQ reagent-labeling performed on the same set of fractions and followed by separate LC MALDI MS/MS analysis produced very similar elution profiles for components of the pyruvate dehydrogenase complex. In both experiments, subunits AceE, AceF, and LpdA had similar elution profiles between fractions E4 and E8, suggesting that sample preparation was fairly reproducible (Figure 5).

To establish the reproducibility of chromatographic separations, two protein fractionation experiments were compared. Both used gel filtration followed by anion exchange chromatography of an *E. coli* lysate but were carried out at different scales using differing amounts of crude extract and different size columns. Despite these significant differences, the RNA polymerase, pyruvate dehydrogenase, and 2-oxoglutarate dehydrogenase complexes all eluted in the same order and maintained very similar elution patterns (Figure 6, panels A and B). Thus, this result demonstrated that it should be possible to compare protein elution profiles between parallel columns at the same step of a tagless fractionation of a single extract or between equivalent columns in different tagless fractionations of extracts derived, for instance, from cells grown under dissimilar conditions to detect differences in protein complex composition.

Identifying Known Protein Complexes. Next, it was necessary to more thoroughly test the feasibility of identifying protein complexes using the iTRAQ approach by examining polypeptide elution profiles of members of known protein complexes.

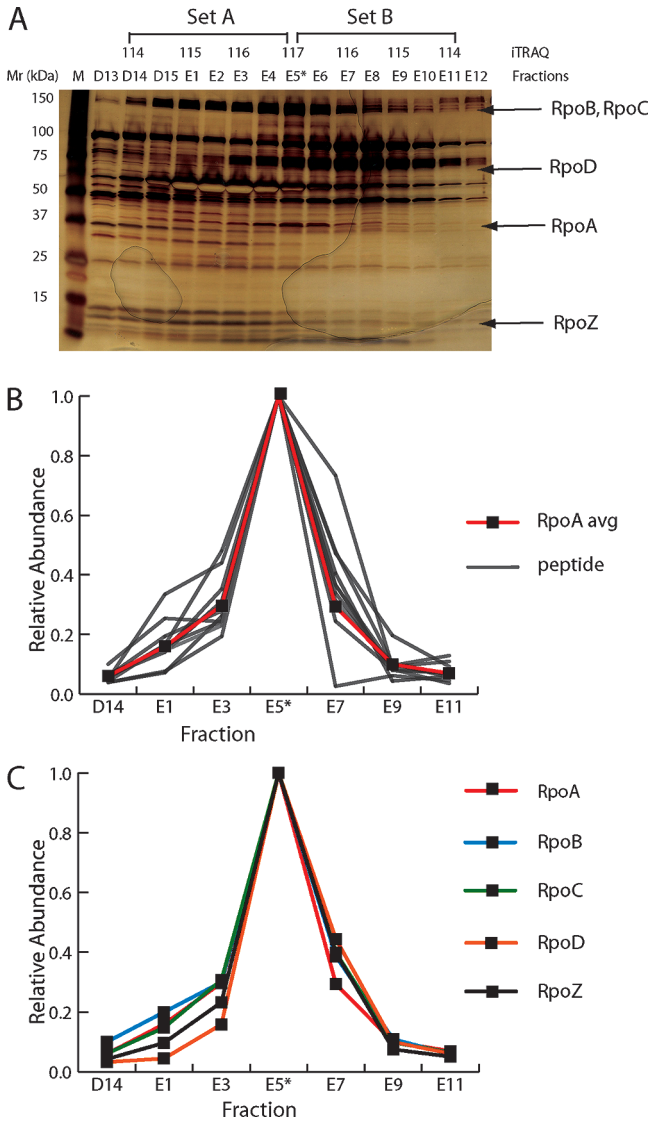

Figure 4. iTRAQ analysis of comigrating RNA polymerase subunits. Panel A shows the result of SDS-PAGE of a subset of fractions from a Mono Q column elute. The positions of components of RNA polymerase complex are annotated with arrows, of which RpoB, -C, and -A are visible within the mixture of other proteins. Two overlapping four-plexes (Set A and Set B) that shared fraction E5 (annotated with an asterisk) were generated by combining every other fraction labeled with an iTRAQ reagent. Panel B shows the relative abundance of each of the different tryptic peptides from RpoA derived after iTRAQ analysis (thin gray lines). The thick red line represents the mean elution profile for RpoA based upon the average for all peptides. Although there is some peptide-to-peptide variation, they all closely approximate the mean. In panel C, the mean elution patterns for the five major polypeptide components of RNAP complex are shown. All closely comigrate under the conditions of this experiment, suggesting that iTRAQ-derived mean relative abundances confidently represent protein elution profiles. For all profiles in panels B and C, the fractions with the maximum levels of polypeptides are set to a nominal relative abundance of 1.0.

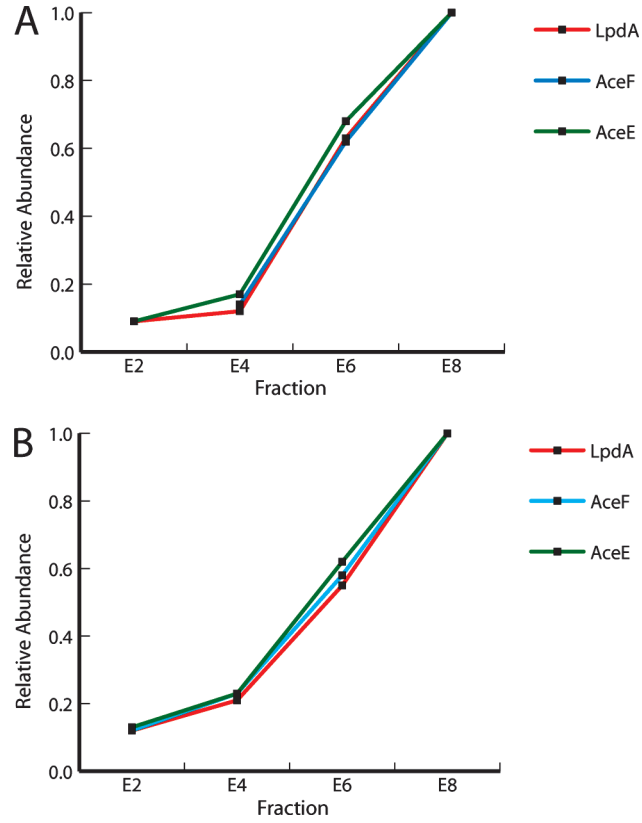

Figure 5. Reproducibility of tryptic digestion, iTRAQ labeling, and LC MALDI MS/MS analysis. A subset of Mono Q column fractions (see panel A in Figure 4) was analyzed in duplicate by independently performing tryptic digestion, iTRAQ labeling, and LC MALDI MS/MS analysis. The mean iTRAQ elution profiles of the three components of pyruvate dehydrogenase complex (LpdA, AceE, AceF) derived from the two experiments are very similar. Panels A and B demonstrate the reproducibility of the methods.

This was accomplished by assaying 15 fractions grouped into five linked multiplex sets from across the larger scale Mono Q chromatography fractions described above. All fractions were initially analyzed utilizing a 4700 Proteomics Analyzer (Applied Biosystems), and then, the first 10 fractions encompassed by three four-plexes were respotted and reanalyzed using 4800 Proteomics Analyzer (Applied Biosystems). A total of 103 nonribosomal polypeptides were identified on the basis of at least one peptide (Supplementary Table A in Supporting Information). Then the literature was consulted to learn how many known protein complexes and protein–protein interactions were to be expected among the polypeptides that were detected. We ignored the fact that for some complexes, usually lower abundance ones, only a subset of polypeptides were identified and instead focused on whether those polypeptides that we could detect were identifiable as comigrating in the iTRAQ data. According to the EcoCyc database (http://biocyc.org/ECOLI/NEW-IMAGE?object=Protein-Complexes), 35 of the polypeptides we detected with the tagless strategy were constituents of 13 known protein complexes, comprising 37 components (one polypeptide, LpdA, participated in 3 protein complexes). According to a large-scale TAP analysis of protein–protein interactions in *E. coli*,[15] 21 of the polypeptides we detected with the tagless strategy were expected to participate in 24 reciprocal pairwise interactions, with 9 polypeptides participating in multiple interactions (Table 1). There was a significant overlap between these two sets of polypeptides since many of the components of known protein complexes in EcoCyc were also detected by TAP methodology. Overall, there was previous evidence that 44 nonribosomal polypeptides

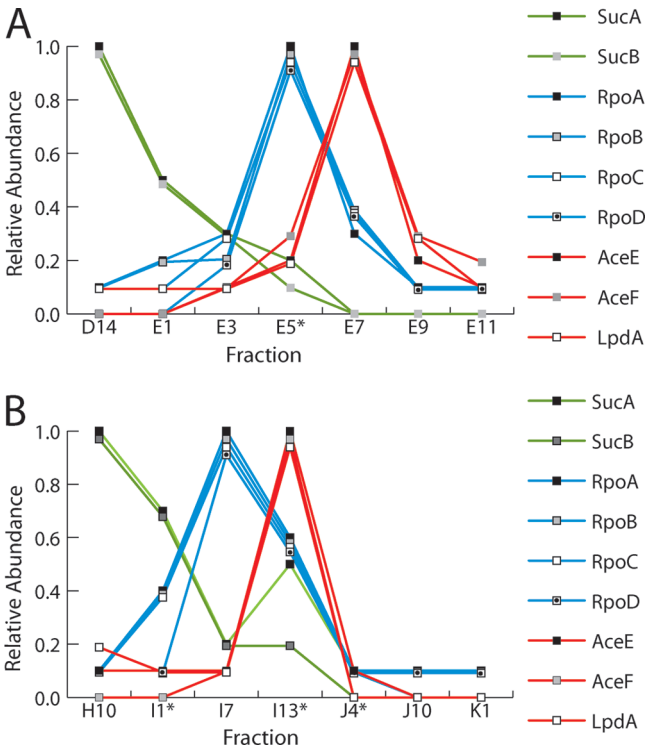

Figure 6. Reproducibility of chromatographic separation. Mean iTRAQ elution profiles of the polypeptide components of RNA polymerase (RpoA, RpoB, RpoC, RpoD), pyruvate dehydrogenase (LpdA, AceE, AceF), and 2-oxoglutarate (SucA, SucB, LpdA) during anion exchange chromatography in two independent protein separation experiments. Panel A shows the results from the same Mono Q fractionation shown in Figures 4 and 5. Panel B show the results from a larger scale Mono Q separation of a different crude extract preparation. The order of elution of complexes in the two different Mono Q experiments is the same, testifying to the feasibility of comparing results between parallel columns at the same step of a tagless fractionation of a single extract or between equivalent column fractionation of two closely related protein extracts.

Table 1. Overview of the Data Set of Identified Nonribosomal Polypeptides

category	number	percentage
All nonribosomal polypeptides	103	100.0%
Polypeptides participating in known protein complexes for which at least two components were detected[a]	35[b]	34.0%
Polypeptides participating in known reciprocal pairwise interactions[c]	21[d]	20.4%

[a] Protein complex data is based upon the content of the Encyclopedia of *E. coli* K-12 Genes and Metabolism (http://biocyc.org/ECOLI/NEW-IMAGE?object=Protein-Complexes). [b] Out of 35 polypeptides, 1 polypeptide (LpdA) participated in three complexes. [c] Protein–protein interaction is data based upon the study of Butland et al.[15] [d] Out of 21 polypeptides, 9 participated in multiple pairwise reciprocal interactions.

(42.7% of the total we identified) were engaged in heteromeric protein–protein interactions.

We first used two *ad hoc* approaches to classify comigrating polypeptides in the iTRAQ data. One was based on polypeptides that showed maximum concentrations in the same fraction (elution apexes). Of the known complex components from the EcoCyc database, the great majority (78.4%) shared the same elution apex (Tables 2 and 3). The second approach scored comigration of polypeptides found in neighboring fractions.

Table 2. Detection of the Expected Protein Complexes and Reciprocal Pairwise Interactions

elution profile characteristics	protein complex components[a] Total = 37[c]	pairwise interactions[b] Total = 24
Apex shared[d]	78.4%	62.5%
Coelution[e]	94.6%	87.5%
No coelution[f]	5.4%	12.5%
Clustered[g]	56.8%	54.2%

[a] Protein complex data are based upon the content of the Encyclopedia of *E. coli* K-12 Genes and Metabolism (http://biocyc.org/ECOLI/NEW-IMAGE?object=Protein-Complexes). [b] Protein–protein interaction data are based upon the study of Butland et al.[15] [c] Three complexes shared the same polypeptide (LpdA), and hence, a number of complex components (37) is higher than a number of identified polypeptides (35). [d] At least two complex components/both interacting partners shared at least one apex of elution. [e] At least two complex components/both interactive partners eluted in the same multiplex, i.e., in neighboring fractions. [f] Complex components/both interacting partners demonstrated apexes of elution in distant fractions. [g] Elution profiles were compared using a modified Pearson's algorithm and clusters were defined employing a threshold of 0.92.

An additional 16.2% were detected in this way, and hence, ~95% of the expected protein complex components demonstrated close coelution (Tables 2 and 3; Supplementary Figure 2 in Supporting Information). Only DNA gyrase components GyrA and GyrB showed completely disparate elution profiles (Supplementary Figure 2A in Supporting Information), but this is not surprising as DNA gyrase is known to be unstable.[33] Of the pairwise interactions verified by reciprocal TAP, 62.5% of partners shared elution apexes and an additional 25% were found in the neighboring fraction, bringing the detection of closely coeluting partners to ~88% (Table 2 and Supplementary Table B in Supporting Information). The iTRAQ approach is even able to identify distinct complexes that share a polypeptide subunit as long as the complexes separate from each other during chromatography. For example, pyruvate dehydrogenase and 2-oxoglutarate dehydrogenase complexes share the LpdA polypeptide in common (Supplementary Figure 2 Panels I and J in Supporting Information). Thus, the iTRAQ strategy does allow a range of different complexes to be detected.

Despite the broad similarity in elution profiles for known complex components, some intriguing differences between their profiles were seen in a few cases. For example, RNA polymerase components NusA and Rho demonstrated much narrower elution peaks than the core RNA polymerase subunits RpoA, -B, and -C (Figure 7A), which likely is due to the different chromatographic properties of the known distinct forms of RNA polymerase causing the core subunits that participate in all forms of polymerase to have broad profiles while particular NusA and/or Rho containing subform(s) instead fractionate more discretely. Thus, even if an agreement between the anticipated and observed elution profiles of known complex components is not complete, it cannot be assumed that this represents an artifact of the tagless methodology or some error in our methods. Rather, an observed discrepancy may reflect biologically relevant differences.

Discovery of Complexes by Automated Cluster Analysis. The above analyses used *ad hoc* criteria to judge if polypeptide elution profiles were sufficiently similar to suggest that they are members of the same protein complex. However, adaptation of the tagless strategy in a high-throughput modality which generates many more fractions will require automated statistical analyses that can identify putative protein complexes and provide confidence estimates on the likelihood of that prediction. Toward accomplishing this goal, a prototype algorithm

Table 3. Components of Known *E. coli* Complexes Identified by the Tagless Strategy

complex ID[a]	polypeptide ID no.	polypeptide code name	uniprot accession no.	ID category[b]	polypeptide MW (Da)	sequence coverage (%)[c]	clustered[d] Yes (1) or No (0)	apex shared[e] Yes (1) or No (0)	coelution[f] Yes (1) or No (0)
A	37	GyrA	P0AES4	[ID2+]	96964	25.4	0	0	0
A	38	GyrB	P0AES6	[ID2+]	89950	5.1	0	0	0
B	54	Lpp	P69776	[ID1]	8323	15.4	1	1	1
B	62	PaL	P0A912	[ID2+]	18824	21.4	1	1	1
B	91	TolB	P0A855	[ID2+]	45956	4.4	1	1	1
C	15	DnaJ	P08622	[ID1]	41044	2.7	0	1	1
C	16	DnaK	P0A6Y8	[ID2+]	69115	37.0	0	1	1
C	34	GrpE	P09372	[ID2+]	21798	25.4	0	1	1
D	6	AtpA	P0ABB0	[ID2+]	55222	1.8	1	1	1
D	7	AtpD	P0ABB4	[ID2+]	50325	6.5	1	1	1
E	14	DnaE	P10443	[ID2+]	129905	1.2	1	1	1
E	17	DnaX	P06710	[ID2+]	71138	4.2	1	1	1
E	43	HolE	P0ABS9	[ID1]	8846	9.2	1	1	1
F	11	CysI	P17846	[ID2+]	63998	4.2	0	1	1
F	12	CysJ	P38038	[ID2+]	66270	8.5	0	1	1
G	61	NusA	P0AFF6	[ID2+]	54871	16.8	1	1	1
G	75	Rho	P0AG30	[ID2+]	47004	18.6	0	0	1
G	79	RpoA	P0A7Z4	[ID2+]	36512	27.7	1	1	1
G	80	RpoB	P0A8V2	[ID2+]	150632	25.3	1	1	1
G	81	RpoC	P0A8T7	[ID2+]	155160	28.4	1	1	1
G	82	RpoD	P00579	[ID2+]	70263	6.2	1	1	1
G	83	RpoZ	P0A800	[ID2+]	10237	44.0	0	0	1
H	58	MukB	P22523	[ID2+]	170230	25.2	0	1	1
H	59	MukE	P22524	[ID2+]	28178	8.6	1	1	1
H	60	MukF	P60293	[ID2+]	50597	9.3	1	1	1
I	1	AceE	P0AFG8	[ID2+]	99668	44.2	1	1	1
I	2	AceF	P06959	[ID2+]	66096	47.8	1	1	1
I	53	LpdA	P0A9P0	[ID2+]	50688	38.0	0	1	1
J	87	SucA	P0AFG3	[ID2+]	105062	25.2	1	1	1
J	88	SucB	P07016	[ID2+]	44011	39.5	1	1	1
J	53	LpdA	P0A9P0	[ID2+]	50688	38.0	0	1	1
K	25	GcvP	P33195	[ID1]	104376	0.9	0	0	1
K	53	LpdA	P0A9P0	[ID2+]	50688	38.0	0	0	1
L	72	PyrB	P0A786	[ID1]	34427	3.2	1	1	1
L	73	PyrI	P0A7F3	[ID1]	17121	6.5	1	1	1
M	31	GlyQ	P00960	[ID1]	44716	3.3	0	0	1
M	32	GlyS	P00961	[ID2+]	76813	23.4	0	0	1

[a] Protein complex data based upon the content of the Encyclopedia of *E. coli* K-12 Genes and Metabolism (http://biocyc.org/ECOLI/NEW-IMAGE?object=Protein-Complexes). [b] Protein identification based on one and two or more peptides for [ID1] and [ID2+] categories, respectively. MS/MS spectra of polypeptides matched to a single peptide are shown in Supplementary Figure 4 in Supporting Information. [c] Unique nonoverlapping peptides were used to calculate protein sequence coverage defined as the ratio between the sum of amino acids encompassed by the confidently matched peptides (%CI >95%) and the number of amino acids in a polypeptide sequence. For polypeptides observed in more than one four-plex, the best four-plex data are shown. [d] At least two complex components were found in the same cluster (modified Pearson's algorithm, a threshold of 0.92): annotation "1" means "Yes", annotation "0" means "No". [e] At least two complex components shared at least one apex of elution: annotation "1" means "Yes", annotation "0" means "No". [f] At least two complex components eluted in the same four-plex, i.e., in neighboring fractions: annotation "1" means "Yes", annotation "0" means "No".

for automatically detecting complexes based on the clustering methods used to detect coregulated genes in expression microarray data[34] was tested.

In general, the elution profile of a polypeptide can be plotted as an intensity map in a multiparameter grid space, where the coordinates of each grid specify a fraction and its intensity indicates the relative abundance of the polypeptide. For example, in a two-step protein complex separation scheme, the map could be plotted exactly like a 3-D geological map representing hills and mountains. The task of finding comigrating polypeptides is then reduced to colocalizing "hill and mountain" peaks within a grid of the N-dimensional map. From the data analysis point of view, each peak is a subset of registered data points and detecting colocalized peaks can be achieved by performing clustering analysis of the subset over the entire collection of protein elution profiles.

Clustering analysis of the whole data set of detected nonribosomal proteins resulted in generation of 17 clusters that grouped 77 polypeptides; no partners were found for the remaining 26 polypeptides (Supplementary Table C in Supporting Information). These results were compared to the manually curated groupings of comigrating polypeptides. Our current clustering algorithm correctly grouped 72.4% and 86.7% of the polypeptides sharing the same apex of elution that were manually classified as members of either EcoCyc known complexes or reciprocal TAP-defined interactions, respectively. The differences between the *ad hoc* and computational methods of grouping polypeptides reflected differences in the criteria used. The manual evaluation was based solely on close comigration of polypeptides defined by elution apexes occurring in the same or neighboring fraction, whereas the clustering algorithm also took into account additional features such as

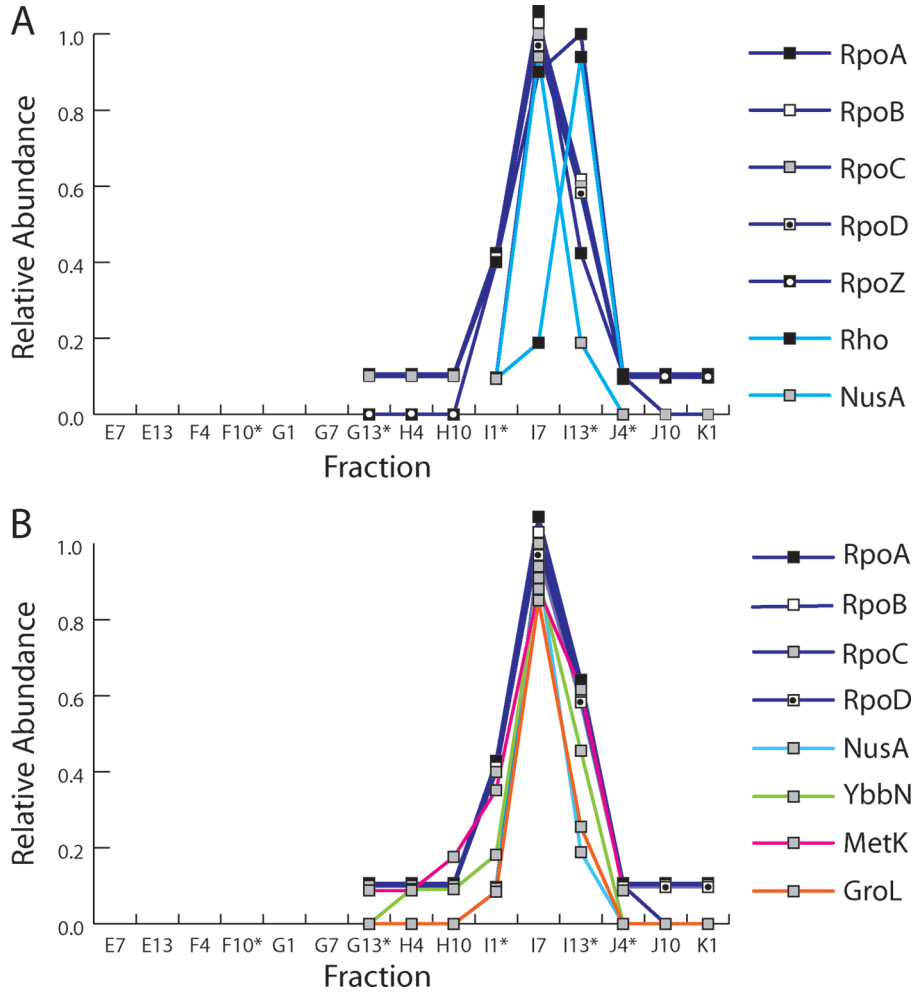

Figure 7. The effectiveness of clustering in grouping complex components. Panel A shows the mean iTRAQ profiles of the five main components of RNA polymerase (RpoA, RpoB, RpoC, RpoD, RpoZ) and also the transcription termination/antitermination factor NusA and transcription termination factor Rho. Panel B shows the constituents of a cluster defined by our algorithm that includes four of the five main RNA polymerase subunits and Nus A, but not RpoZ and Rho. The cluster also contained proteins that had likely fortuitously coeluted with RNA polymerase: YbbN, MetK, GroL, and AldA. We note that there is strong TAP evidence that GroL and YbbN might form significant interaction.

peak shape, peak resolution and the presence of multiple apexes within a contiguous portion of polypeptide elution profile. These additional constraints resulted in the exclusion of some of the known complex components from certain clusters. For example, LpdA demonstrated multiple apexes of elution that overlapped with pyruvate dehydrogenase and 2-oxoglutarate complex components but clustered with neither (Supplementary Table C and Figure 2, Panels I and J, in Supporting Information). The RNA polymerase components RpoA, RpoB, RpoC, RpoD, and NusA were included in the same cluster, but RpoZ and Rho were not (Figure 7B). Likewise, two components of a chromosome partitioning complex, MukE and MukF, were placed in the same cluster, but another component, MukB, was not because it eluted in a narrower peak (Supplementary Table 3 and Figure 1, Panel H, in Supporting Information). The "failure" to cluster all complex components as one entity might actually provide important insights into the previously discussed heterogeneity between distinctly eluting "subcomplexes". Therefore, the future development of the clustering algorithm will neccessitate the incorporation of more sophisticated tools, for example, flexible, multitiered stringency scales to allow more robust recognition of potential coeluting polypeptides while preserving detailed information about chromatographic peak shape and resolution.

The data inputted into the clustering analysis was not limited to members of known complexes, but included all 103 nonribosomal proteins. Not surprisingly, given the crude and complex nature of the chromatography fractions analyzed, clusters that included members of known complexes frequently contained additional polypeptides that seemed unlikely to be uncharacterized members of these complexes. For example, YbbN, MetK, and GroEL clustered with members of the RNA polymerase complex (Figure 7B), but have not been associated with this well-studied complex before. At the same time, TAP data strongly point to the possibility of a significant [YbbN: GroL] interaction.[15] We suspect that the majority of these additional cluster members resulted from fortuitous coelution of either single polypeptides or components of (multiple) distinct protein complexes within the Mono Q column chromatogram. Incorporation of additional protein separation steps into a full scale tagless separation scheme should greatly reduce the frequency of such "opportunistic coeluting", especially since clustering can then take into account comigration across not one but three of four dimensions. With much more

information available, it should also be possible to better distinguish between potential subcomplex forms and identify novel complexes and complex components.

Purification of Protein Complexes. The tagless strategy was able to identify putative protein complexes without a need for complete purification. However, these results indicate that a majority of high and moderately abundant stable protein complexes can be purified to near homogeneity by an optimized tagless fractionation method employing four orthogonal separation steps and scaling up the amount of starting material. Even with the pilot fractionations employed here, that is, using only anion exchange and size exclusion chromatography, three complexes (pyruvate hydrogenase, RNA polymerase, and GroEL) have been purified to apparent homogeneity from *E. coli* cell lysate (Supplementary Figure 3 in Supporting Information). More extensive fractionation of extracts from the sulfate reducing bacteria *Desulfovibrio vulgaris* has, to date, purified 3 heteromeric and 25 homomeric complexes from only a small portion of the total fractionation space.[35] These purified complexes are amenable to further characterization methods such as single particle electron microscopy and Small Angle X-ray Scattering. For example, the 17 Å structure of one complex identified and purified by the tagless scheme has already been obtained.[19]

Conclusions

We have established proof-of-principle evidence for the feasibility of employing a tagless strategy for protein complex identification and purification. We estimate that at least around 50% of bacterial polypeptides participate in complexes that are sufficiently stable to survive the multiple chromatographic steps. LC MALDI-based tagless strategy is blind to complex stoichiometry—as long as polypeptide components are detected, their elution profiles will be drawn and the presence of complexes inferred. Hence, we expect that both stoichiometric and substoichiometric complexes will be identified. The range of complexes identified by the tagless strategy is likely to be comparable to those identified by TAP experiments. Out of 24 TAP-detected reciprocal interactions,[15] only three (MetK−SecA, MetK−DnaJ, and GyrA−GyrB) had completely dissociated during purification (Supplementary Table B in Supporting Information). In addition, there is good reason to believe that those complexes that are disrupted by the use of an affinity tag and, therefore, are not detectable by TAP, will be identifiable by a tagless approach. Relative quantification using iTRAQ reagents allowed comigration of polypeptides to be determined and the chromatographic separation appeared sufficiently reproducible such that results across multiple parallel chromatograph columns, each separating different subsets of total cellular protein, could be meaningfully compared. Even a relatively simple clustering algorithm was effective at automatically detecting members of protein complexes using data from only two dimensions of separation. Several of the more abundant complexes were purified to greater than 70% homogeneity.

Although these results are encouraging, the pilot scale fractionation and mass spectrometry analysis described above identified only 103 nonribosomal polypeptides and 13 protein complexes, whereas there are approximately 3000 known water soluble polypeptides participating in several hundred stable complexes in *E. coli*. What further improvements will be needed to detect all proteins and stable protein complexes in a realistic time scale? First, the samples analyzed by iTRAQ LC MALDI MS/MS were derived from a subset of the protein fractions of a two-dimensional scheme and represented only approximately 10%, by mass, of the water-soluble *E. coli* proteins. Thus, even at this current pilot scale, it is likely that around 1000 polypeptides would have been detected had all fractions from the scheme been analyzed by mass spectrometry. The remaining 2000 or so water-soluble proteins that would not have been detected are in most cases likely to be of lower abundance. Hence, by starting with a large amount of crude protein extract and employing four, rather than two, orthogonal chromatography separation steps, it should be possible to detect the great majority of these lower abundance polypeptides. Of the two constraints inherent to analysis of low-abundance species, that is, dynamic range challenges and availability of material, the former is currently being addressed by performing extensive protein separation involving multiple chromatographic steps. The latter constraint is not a major obstacle since biomass for our target organism *D. vulgaris* is currently produced on a 400 L scale (4×10^{13} to 4×10^{14} cells) that delivers ~10 g soluble protein (~ 200 μmol of total protein, assuming an average polypeptide MW of 50 kDa). Within this mixture, a low-abundance polypeptide expressed at the level of 10 copies per cell will constitute ~670 pmol material that corresponds to a 3.3×10^{-6} portion of total protein. The current yield after the four protein complex separation steps, tryptic digestion, and iTRAQ-labeling is estimated at ~0.5%. Assuming the same level of recovery of low-abundance complex components and anticipating a spread of protein complex elution during a 4-step fractionation into 50 fractions, 3.35 pmol of the low-abundance protein will be recovered at a level of 67 fmol per fraction or ~130 fmol per iTRAQ multiplex, assuming the worst case situation when only two fractions within a four-plex might contain a protein complex. This scenario brings us within the current practical detection limits of the MALDI TOF/TOF instrument. With an expected increase in the sensitivity of mass spectrometers over the next 5−10 years, nearly all complexes should be detectable with such a fractionation. We have now established a four-dimensional fractionation at this lager scale and are now optimizing each fractionation step (unpublished data). While the success of discovery of any specific low level protein complex will be highly dependent on the extent of its separation from other species, efficiency of digestion and labeling, and quality of MS/MS, in principle, detection of low-abundance complexes is within the realm of possibility.

Another improvement needed is to automate and speed up many steps. A major advantage of the tagless approach is that by its design it is intrinsically more amenable to automation than TAP as it consists of fewer types of operations and is highly repetitious. For example, no genetic manipulation of the organism is required and only one large culture of cells need be grown. With the automation of the sample preparation and chromatographic separations and development of a data analysis pipeline that is coupled to real time control of the mass spectrometer to eliminate redundant and time-consuming analysis of peptides from the same protein, and the expected future increase in the speed of MALDI MS/MS instruments, it should be possible to achieve much higher throughput identification of protein complexes than is currently possible.

Finally, additional methods to establish the accuracy and veracity of putative complexes identified by the tagless strategy will be needed. We certainly expect an increase in the number of fractionation steps and the use of more complex clustering algorithms that employ quantitative data on the migration of

polypeptides across four chromatographic dimensions to reduce the occurrences of "opportunistic coeluting" of unrelated proteins seen in the pilot study. While clustering data from a single protein complex separation step is not capable of discerning 'true' complexes from 'fortuitous coeluting polypeptides', protein clustering within a multidimensional separation space should be capable of detecting discrepancies in elution profiles of putative complex candidates and thus eliminate many cases of fortuitous coelution. At least in some model organism a subset of putative complexes could and should be verified by reciprocal TAP analysis. In general, it is critical to cross-verify the predictions made by any method to identify protein complexes system-wide using a combination of biological and analytical techniques.

In conclusion, the tagless protein complex identification strategy is a discovery as well as a purification tool. Its great strengths lie in the ability to analyze native systems and in the potential of highly automated high-throughput execution. We expect that a combination of tagless- and immunoaffinity-based complex isolation strategies will greatly expand the amount of information about the biology of organisms and provide orthogonal confirmation of the overlapping results.

Acknowledgment. We are grateful to all members of the Protein Complex Analysis Project (PCAP) team (http://vimss.lbl.gov/projects/pcap.html) for their support and discussion and especially to Terry Hazen for his unwavering drive to facilitate publication of this paper. We thank Jeremy Semeiks and Gavin Sherlock for supplying the source code of the "xcluster" software, and useful discussions regarding the implementation of the clustering algorithm. Scott Dixon is acknowledged for his excellent technical assistance in all aspects of mass spectrometric analysis. The UCSF Mass Spectrometry Core Facility is supported in part by the Sandler Family Foundation. This work was conducted under Department of Energy contract DE-AC02-05CH11231 awarded to Lawrence Berkeley National Laboratory.

Supporting Information Available: Tables, Supplementary Table A, nonribosomal proteins identified by tagless strategy; Supplementary Table B, tagless strategy-detection of reciprocal protein-protein interactions previously identified by TAP; Supplementary Table C, results of clustering analysis of elution profiles of nonribosomal proteins. Figures, Supplementary Figure 1, generation of polypeptide elution profiles; Supplementary Figure 2, Elution profiles of detected components of known protein complexes (panels A−M); Supplementary Figure 3, SDS-PAGE of 3 protein complexes purified from *E. coli* lysate using tagless strategy; Supplementary Figure 4, evidence of identification of polypeptides matched to a single peptide, annotated MS/MS spectra ([ID1] category, 27 panels. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Buchanan, M. V.; Larimer, F. W.; Wiley, H. S.; Kennel, S. J.; Squier, T. J.; Ramsey, J. M.; Rodland, K. D.; Hurst, G. B.; Smith, R. D.; Xu, Y.; Dixon, D.; Doktycz, M. J.; Colson, S.; Gesteland, R.; Giometti, C.; Young, M.; Giddings, M. Genomes to Life "Center for Molecular and Cellular Systems": a research program for identification and characterization of protein complexes. *OMICS* **2002**, *6*, 287–303.

(2) McHenry, C. S.; Crow, W. DNA polymerase III of Escherichia coli. Purification and identification of subunits. *J. Biol. Chem.* **1979**, *254*, 1748–1753.

(3) Srere, P. A.; Mathews, C. K. In *Guide to Protein Purification*; Deutscher, M. P., Ed.; Academic Press: San Diego, CA, 1990; Vol. 182, pp 539–551.

(4) Austin, R. J.; Biggin, M. D. Purification of the Drosophila RNA polymerase II general transcription factors. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5788–5792.

(5) Link, A. J.; Fleischer, T. C.; Weaver, C. M.; Gerbasi, V. R.; Jennings, J. L. Purifying protein complexes for mass spectrometry: applications to protein translation. *Methods* **2005**, *35*, 274–290.

(6) Balbo, A.; Minor, K. H.; Velikovsky, C. A.; Mariuzza, R. A.; Peterson, C. B.; Schuck, P. Studying multiprotein complexes by multisignal sedimentation velocity analytical ultracentrifugation. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 81–86.

(7) Camacho-Carvajal, M. M.; Wollscheid, B.; Aebersold, R.; Steimle, V.; Schamel, W. W. Two-dimensional Blue native/SDS gel electrophoresis of multi-protein complexes from whole cellular lysates: a proteomics approach. *Mol. Cell. Proteomics* **2004**, *3*, 176–182.

(8) Rout, M. P.; Aitchison, J. D.; Suprapto, A.; Hjertaas, K.; Zhao, Y.; Chait, B. T. The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **2000**, *148*, 635–651.

(9) Rigaut, G.; Shevchenko, A.; Rutz, B.; Wilm, M.; Mann, M.; Seraphin, B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **1999**, *17*, 1030–1032.

(10) Puig, O.; Caspary, F.; Rigaut, G.; Rutz, B.; Bouveret, E.; Bragado-Nilsson, E.; Wilm, M.; Seraphin, B. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **2001**, *24*, 218–229.

(11) Winkler, G. S.; Lacomis, L.; Philip, J.; Erdjument-Bromage, H.; Svejstrup, J. Q.; Tempst, P. Isolation and mass spectrometry of transcription factor complexes. *Methods* **2002**, *26*, 260–269.

(12) Burckstummer, T.; Bennett, K. L.; Preradovic, A.; Schutze, G.; Hantschel, O.; Superti-Furga, G.; Bauch, A. An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat. Methods* **2006**, *3*, 1013–1019.

(13) Gavin, A. C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L. J.; Bastuck, S.; Dumpelfeld, B.; Edelmann, A.; Heurtier, M. A.; Hoffman, V.; Hoefert, C.; Klein, K.; Hudak, M.; Michon, A. M.; Schelder, M.; Schirle, M.; Remor, M.; Rudi, T.; Hooper, S.; Bauer, A.; Bouwmeester, T.; Casari, G.; Drewes, G.; Neubauer, G.; Rick, J. M.; Kuster, B.; Bork, P.; Russell, R. B.; Superti-Furga, G. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **2006**, *440*, 631–636.

(14) Krogan, N. J.; Cagney, G.; Yu, H.; Zhong, G.; Guo, X.; Ignatchenko, A.; Li, J.; Pu, S.; Datta, N.; Tikuisis, A. P.; Punna, T.; Peregrin-Alvarez, J. M.; Shales, M.; Zhang, X.; Davey, M.; Robinson, M. D.; Paccanaro, A.; Bray, J. E.; Sheung, A.; Beattie, B.; Richards, D. P.; Canadien, V.; Lalev, A.; Mena, F.; Wong, P.; Starostine, A.; Canete, M. M.; Vlasblom, J.; Wu, S.; Orsi, C.; Collins, S. R.; Chandran, S.; Haw, R.; Rilstone, J. J.; Gandi, K.; Thompson, N. J.; Musso, G.; St Onge, P.; Ghanny, S.; Lam, M. H.; Butland, G.; Altaf-Ul, A. M.; Kanaya, S.; Shilatifard, A.; O'Shea, E.; Weissman, J. S.; Ingles, C. J.; Hughes, T. R.; Parkinson, J.; Gerstein, M.; Wodak, S. J.; Emili, A.; Greenblatt, J. F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **2006**, *440*, 637–643.

(15) Butland, G.; Peregrin-Alvarez, J. M.; Li, J.; Yang, W.; Yang, X.; Canadien, V.; Starostine, A.; Richards, D.; Beattie, B.; Krogan, N.; Davey, M.; Parkinson, J.; Greenblatt, J.; Emili, A. Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **2005**, *433*, 531–537.

(16) Dong, M.; Biggin, M. D.; Williams, K.; Dixon, S. E.; Yang, L. L.; Fisher, S. J.; Hall, C. S.; Jin, J.; Witkowska, H. E. Multi-dimensional orthogonal separation and iTRAQ™ reagent tracking: a genome-wide "tagless" strategy for isolation and detection of soluble bacterial protein complexes. Presented at the 54th ASMS Conference on Mass Spectrometry and Allied Techniques, Seattle, WA, 2006.

(17) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, *3*, 1154–1169.

(18) Andersen, J. S.; Wilkinson, C. J.; Mayor, T.; Mortensen, P.; Nigg, E. A.; Mann, M. Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **2003**, *426*, 570–574.

(19) Garczarek, F.; Dong, M.; Typke, D.; Witkowska, H. E.; Hazen, T. C.; Nogales, E.; Biggin, M. D.; Glaeser, R. M. Octomeric pyruvate-

ferredoxin oxidoreductase from Desulfovibrio vulgaris. *J. Struct. Biol.* **2007**, *159*, 9–18.
(20) Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **1976**, *72*, 248–254.
(21) Medzihradszky, K. F.; Campbell, J. M.; Baldwin, M. A.; Falick, A. M.; Juhasz, P.; Vestal, M. L.; Burlingame, A. L. The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal. Chem.* **2000**, *72*, 552–558.
(22) Keseler, I. M.; Collado-Vides, J.; Gama-Castro, S.; Ingraham, J.; Paley, S.; Paulsen, I. T.; Peralta-Gil, M.; Karp, P. D. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res.* **2005**, *33*, D334–337.
(23) Zieske, L. R. A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies. *J. Exp. Bot.* **2006**, *57*, 1501–1508.
(24) Bondarenko, P. V.; Chelius, D.; Shaler, T. A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal. Chem.* **2002**, *74*, 4741–4749.
(25) Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **2003**, *75*, 4818–4826.
(26) Liu, H.; Sadygov, R. G.; Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **2004**, *76*, 4193–4201.
(27) Zhang, B.; VerBerkmoes, N. C.; Langston, M. A.; Uberbacher, E.; Hettich, R. L.; Samatova, N. F. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.* **2006**, *5*, 2909–2918.
(28) Bublitz, R.; Kreusch, S.; Ditze, G.; Schulze, M.; Cumme, G. A.; Fischer, C.; Winter, A.; Hoppe, H.; Rhode, H. Robust protein quantitation in chromatographic fractions using MALDI-MS of tryptic peptides. *Proteomics* **2006**, *6*, 3909–3917.
(29) Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **2007**, *389*, 1017–1031.
(30) Hartman, N. T.; Sicilia, F.; Lilley, K. S.; Dupree, P. Proteomic complex detection using sedimentation. *Anal. Chem.* **2007**, *79*, 2078–2083.
(31) Dunkley, T. P.; Watson, R.; Griffin, J. L.; Dupree, P.; Lilley, K. S. Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* **2004**, *3*, 1128–1134.
(32) Sadowski, P. G.; Dunkley, T. P.; Shadforth, I. P.; Dupree, P.; Bessant, C.; Griffin, J. L.; Lilley, K. S. Quantitative proteomic approach to study subcellular localization of membrane proteins. *Nat. Protocols* **2006**, *1*, 1778–1789.
(33) Higgins, N. P.; Peebles, C. L.; Sugino, A.; Cozzarelli, N. R. Purification of subunits of Escherichia coli DNA gyrase and reconstitution of enzymatic activity. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 1773–1777.
(34) Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14863–14868.
(35) Dong, M.; Liu, H.; Allen, S.; Hall, S.; Fisher, S.; Hazen, T.; Geller, J.; Singer, M.; Yang, L.; Jin, J.; Biggin, M.; Witkowska, H. E. Methodological Refinements in iTRAQ™ Reagent-Based "Tagless" Strategy of Identification and Purification of Soluble Protein Complexes in Bacteria. Presented at the 8th International Symposium on Mass Spectrometry in the Health and Life Sciences: Molecular and Cellular Proteomics, San Francisco, CA, 2007.

PR700624E